Visual Programming: Compositional visual reasoning without training

Year: 2022 | Citations: 544 | Authors: Tanmay Gupta, Aniruddha Kembhavi

---

Abstract

We present Visprog, a neuro-symbolic approach to solving complex and compositional visual tasks given natural language instructions. Visprog avoids the need for any task-specific training. Instead, it uses the incontext learning ability of large language models to generate python-like modular programs, which are then executed to get both the solution and a comprehensive and interpretable rationale. Each line of the generated program may invoke one of several off-the-shelf computer vision models, image processing subroutines, or python functions to produce intermediate outputs that may be consumed by subsequent parts of the program. We demonstrate the flexibility of VIsPROG on 4 diverse tasks - compositional visual question answering, zero-shot reasoning on image pairs, factual knowledge object tagging, and language-guided image editing. We believe neuro-symbolic approaches like Visprog are an exciting avenue to easily and effectively expand the scope of AI systems to serve the long tail of complex tasks that people may wish to perform.